Cross-linguistic pragmatic differences as a function of hyponym viability

Danielle Dionne

June 15, 2020

Abstract

The goal of this paper is to investigate what determines the viability of alternatives in pragmatic reasoning about scalar implicatures. Data from English and Spanish speakers is compared to investigate whether cross-linguistic differences in scalar implicature calculation are based on how idiomatic an alternative is. The data supports the idea that viability of an utterance as an alternative is determined based on more than simply complexity. Namely, alternative utterances are viable based on how prevalent, or common, they are. Languages differ in regard to which utterances are considered viable because translational equivalents in some languages are more widespread than others in a different language. When comparing two Rational Speech Act models, a prevalence-based model, which considers production data that I collected, significantly outperforms a complexity-based model that considers length of utterance in words. This suggests that listeners are Bayesian – they are informed by production probabilities, which reflect how widespread an alternative is, when calculating scalar implicatures. Additionally, my findings provide evidence against a structural approach to calculating alternatives (Horn, 2000; Katzir, 2007), favoring theories that determine alternatives based

on production probability (Geurts, 2011; Goodman & Stuhlmüller, 2013).

Contents

1 Introduction

Quantity implicature, also known as scalar implicature[1], is the type of implicature drawn when an utterance contains a weaker form which implies the negation of a stronger alternative on the same scale. Alternative utterances, also referred to simply as alternatives, are relevant utterances that a speaker could have said, but chose not to. For example, in English, when someone says *She ate some of the grapes*, a listener may draw the conclusion that the speaker meant *some and not all* of the grapes. This is because the speaker's decision to use *some* over the stronger alternative *all* implies that *all* is false. One big question surrounding this pragmatic reasoning process is how alternatives are activated. Understanding how alternatives are activated allows for a better understanding of scalar implicature. In order for a scalar implicature to exist, stronger alternatives must be activated upon the use of a weaker term on the same scale. The goal of the present paper is to compare scalar implicature calculation cross-linguistically and determine whether differences in the prevalence of alternatives, or how colloquial an alternative is, affect these scalar implicatures in each language.

Within theories of scalar implicature, theorists have offered different methods for constraining the set of alternatives. Multiple theories surrounding the calculation of alternatives refer directly to the notion of complexity (Frank & Goodman, 2012; Horn, 2000; Katzir, 2007 to name a few). Reexamining the example from above, we see that *some* and *all* are equally complex – they are both one word. Theorists that include complexity might argue that this equal complexity informs the listener's calculation of the implicature *some* $\rightsquigarrow$ 'not all'. Within this line of research, there exists a divide in how large a role complexity plays. Some linguists argue in support of a strong view on the role of complexity:

[1] There are likely quantity implicatures that are not scalar implicatures, but for the purposes of this paper, I am using these terms interchangeably.

that complexity determines whether an utterance is considered as an alternative (Katzir, 2007; Horn, 1984; and Horn, 2000). Other linguists have argued for a weak role of complexity, saying that complexity has a small role in determining alternatives (Swanson, 2010). A third group of linguists situate themselves somewhere in the middle, where complexity plays a role in determining alternatives, but in conjunction with how prevalent the alternative is (Goodman & Stuhlmüller, 2013 & Geurts, 2011). While the role of complexity in constraining the set of alternatives has been thoroughly investigated, little research has been conducted, however, that examines the role of prevalence – i.e. how common it is to use a particular utterance to refer to a particular entity– in scalar implicature. Even further, no research, to my knowledge, has been conducted that examines the role of prevalence through cross-linguistic comparison. What, for example, might happen in a language whose translational equivalent of *all* is less colloquial than English *all*?

One example that is especially good for cross-linguistic comparison is the implicature *finger* $\rightsquigarrow$ 'not thumb', since digits are concrete and accessible across languages. Horn (1984) and Geurts (2011) have both invoked complexity in their theories of scalar implicature, and have made specific mention of the implicature *finger* $\rightsquigarrow$ 'not thumb'. Horn (2000) elucidates the *finger* and *toe* contrast, implying a series of specific predictions surrounding scalar implicature. He speculates that if the colloquial language replaced *thumb* with the term *pollex* (the English scientific term for the thumb without the hand/foot contrast) the asymmetry would disappear. Geurts (2011) remarks on Horn's (1984) suggestion, emphasizing the importance of the term *colloquial*. Crucially, it is this "colloquiality" that will be under investigation here.

The present paper analyzes cross-linguistic differences in scalar implicature calculation as a function of the prevalence of alternatives. The next section outlines multiple frameworks for scalar implicature, the constraints these frameworks impose on alternatives, and how digits fit into these frameworks. Section three outlines the methodology and results of the production studies. The fourth section lists my research questions. In the fifth section I provide predictions for the comprehension studies. Section six outlines the methodologies and results of the comprehension studies. The seventh section details the two Rational Speech Act models, and the eighth section discusses the implications of the empirical and model results. Finally, in the ninth section I draw conclusions based on my findings, outline new questions, and describe areas of future research.

2 Background

In order to frame the current study within existing literature, I will begin by outlining multiple frameworks of scalar implicature. In the second subsection, I describe the different methods some of these frameworks possess for determining the set of alternatives. In the third subsection, I introduce the concrete examples that I use as the basis for my experiments – *finger* and *toe*.

2.1 Frameworks for Scalar Implicature

Scalar implicature is the kind of implicature in which a listener assumes that the speaker's use of a weaker term implies the negation of a stronger term. One such example, previously described in the introduction, is the implicature *some* $\rightsquigarrow$ 'not all'. Another example of scalar implicature in English is *rectangle* $\rightsquigarrow$ 'not square' (Horn, 1984). A rectangle is defined as having four sides that meet at right angles. Thus, a square, by definition, still falls under the category of rectangle. However, with scalar implicature, *rectangle* narrows in meaning, in most contexts implying 'not square'.[2] A third example of scalar implicature is *finger* $\rightsquigarrow$ 'not thumb' in English. It is generally agreed upon that humans have ten fingers, so the thumb is still broadly classified as a finger. However, if someone were to say *She has a tattoo on her finger*, the implicature 'not thumb' would be computed, since *thumb* was not uttered.

Pragmatic theorists have offered various ways of spelling out the pragmatic reasoning process behind scalar implicature. Some theories outline the pragmatic reasoning process solely from the hearer's perspective (Grice, 1981; Horn, 1984). More recently, some theorists have presented a pragmatic reasoning process that incorporates a competence assumption, which is to say that the pragmatic reasoning process is based in part on the speaker's belief state (Zimmermann, 2000; van Rooij & Schulz, 2004; Sauerland, 2004; Geurts, 2005, 2011). A third group of theories makes use of the hearer and the speaker perspective in a recursive manner when outlining the pragmatic reasoning process behind scalar implicature (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). There are also theories that argue that scalar implicatures are not pragmatic, but instead are derived within the grammar (Chierchia, 2004a; Chierchia et al., 2008).

Grice (1981) focuses on the perspective of the listener when outlining the pragmatic reasoning process behind conversational implicatures. First, a speaker utters a sentence, ϕ. The speaker's utterance, ϕ, disobeys Grice's Cooperative Principle ("make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged") unless an alternative utterance, ψ, is true (Grice 1975, p. 45). Thus, the listener reasons that ψ must be true. To spell this out in concrete examples, let us consider the following dialogue:

[2] It is important to note that Horn (1984) discusses a process called *autohyponymy*, in which a broader term undergoes Q-based narrowing and becomes a hyponym of itself.

(1) a. How many grapes did she eat?

b. She ate some of the grapes.

c. $\rightsquigarrow$ She did not eat all of the grapes.

Given the goal of the dialogue, the statement *She ate some of the grapes* is less informative than is required to answer the question in (1-a). Based on Gricean pragmatic reasoning, this speech act would disobey the Cooperative Principle unless (1-c) (ψ) is also true (1975). The listener assumes the speaker is obeying the Cooperative Principle and concludes that the speaker believes (1-c) is true. This method of spelling out the pragmatic reasoning process behind scalar implicature is hearer-centric, outlining scalar implicature based on the listener's reasoning.

Horn (1984), a Neo-Gricean theorist, distills Grice's (1975) maxims into two broader principles and uses those principles to explain scalar implicature. He condenses the maxims into the R-principle, "Say no more than you must (given Q)", and the Q-principle, "say as much as you can (given R)" (Horn, 1984, p. 13). Horn (1984) outlines the pragmatic reasoning process behind scalar implicatures from the hearer's perspective, referring to them as Q-based inferences. Given the utterance *She ate some of the grapes*, Horn (1984) argues that the maxims license the listener to draw the conclusion *she ate some but not all of the grapes.* This is under the assumption that the speaker is required to obey the Q-principle "to say as much as possible", which leads the listener to believe that the stronger form *all* would not hold.

One way of outlining the pragmatic reasoning process of scalar implicature that incorporates the speaker's belief state is codified by Geurts (2011) as the "Standard Recipe". Geurts refers to this framework as the "standard recipe" because of its popularity in previous research (see Zimmermann, 2000; van Rooij & Schulz, 2004; Sauerland, 2004; Geurts, 2005). The "Standard Recipe" builds directly on Gricean maxims, but incorporates what van Rooij & Schulz (2004), Sauerland (2004) and others have referred to as a competence or experthood assumption, which means the listener is reasoning about the beliefs of the speaker. According to the "Standard Recipe", a listener first posits, after hearing ϕ, that the speaker could have made a stronger statement: ψ. The listener then assumes that the speaker has an opinion on the truth value of the alternative utterance, ψ. This competence assumption is an additional step that is not present in Grice's (1975) outline of conversational implicature. The listener then draws the conclusion that the alternative utterance, ψ, is not believed to be true. Thus, for the first example mentioned above, uttering *She ate some of the grapes* leads a listener to assume the speaker has an opinion on the stronger alternative. The listener concludes that the speaker believes *She ate all of the grapes* is false, implicating *She did not eat all of the grapes.*

Although the "Standard Recipe" is outlined in Geurts (2011), Geurts actually proposes a framework for scalar implicature that is slightly different. In what he refers to as the "intention-based approach" to scalar implicature, Geurts (2011) begins with intentional states that the speaker could have been in (p. 110). A speaker utters a sentence, ϕ, and the hearer reasons that the speaker believes ϕ. The hearer then considers the set of possible belief states that the speaker could have been in. Finally, the hearer uses alternative utterances to eliminate any belief states that are inconsistent with what the speaker said. Returning to the example *She ate some of the grapes*, Geurts (2011) argues that the hearer will consider belief states such as those listed in the example below.

(2) a. BEL_S(She ate all of the grapes)
    b. $BEL_S(\neg$(She ate all of the grapes))

The hearer then reasons that the (3-a) could have been conveyed more easily if the speaker said *She at all of the grapes.* Assuming that the hearer believes the speaker is competent regarding the proposition that she ate all of the grapes, the hearer would discard (3-a) and conclude that the speaker is in belief state (3-b).

The Rational Speech Act (RSA) framework also outlines the pragmatic reasoning process from both the listener's perspective and the speaker's perspective. The RSA framework uses probabilistic reasoning to model the recursive nature of pragmatic reasoning. It consists of three models that work together to predict listener comprehension and speaker production – a literal listener model, a pragmatic listener model, and a speaker model. According to this framework, a speaker reasons about a listener who reasons about the speaker's knowledge state to determine what utterances will convey the appropriate meaning (Yuan et al., 2018).

$$L_0(s \,|\, u) \propto [\![u]\!](s) \cdot P(s)$$

The literal listener model (above) is the probability that a 'literal listener' will choose state s given utterance u. A literal listener assigns equal probability to every state compatible with the literal meaning, modulo the prior.

Given a state s, the speaker model, S, in the RSA framework chooses an utterance, u, based on accuracy (α) and cost (β). Crucially, the speaker model considers a literal listener's probability of choosing a state, s, given an utterance, u, and the pragmatic listener model considers the speaker's probability of uttering u with the intention of conveying s.

$$S(u \,|\, s) \propto \mathsf{exp}(\alpha \cdot L_0(s \,|\, u) - \beta \cdot \mathsf{length}(u))$$

In RSA, the pragmatic listener (below) chooses an interpretation, s, of an utterance, u, according to the likelihood of the utterance, which is proportional to the speaker's probability of producing u while intending to communicate s, and the prior probability of s.

$$L(s\,|\,u) \propto S(u\,|\,s) \cdot P(s)$$

The RSA framework presents a gradient approach to alternatives in scalar implicature, which is to say that constraints on the set of alternatives do not rule alternatives out entirely, but instead assign such low probabilities that they are never selected (Franke & Jäger, 2016). Additionally, the RSA framework takes a position surrounding the question of whether speakers and listeners are Bayesian. Bayesian pragmatic reasoning follows on the idea that pragmatic theory is probabilistic, rational, and interactive (Franke & Jäger, 2016). A Bayesian listener's interpretation arises after considering the state of the world necessary for the speaker to utter what they did. A speaker's choice of utterance arises after the speaker considers what utterance would be most accurately interpreted by the listener. The recursive, probabilistic nature of RSA suggests the idea that speakers and listeners are Bayesian in nature. This particular approach to pragmatic reasoning is a relatively new one, but the empirical evidence supporting it is substantial (see Bergen et al., 2012; Frank & Goodman, 2012; Bergen et al., 2016; Goodman & Frank, 2016).

In addition to the pragmatic theories outlining scalar implicature that I describe above, there exists a series of theories that promote a grammatical approach to outlining scalar implicatures (Chierchia, 2004a, 2005; Fox, 2007; Chierchia et al., 2008 and others). Chierchia (2004a) outlines a framework for scalar implicature that stems from the lexicon. He proposes that scalar terms, such as *some*, possess two semantic values: one that is "plain" – free of scalar implicature – and one that is "strengthened" – possessing a scalar meaning (p. 59). Instead of a listener calculating a scalar implicature due to a metalinguistic awareness of alternative utterances that were not said, the speaker utters a sentence using the strengthened meaning and that strengthened meaning can be cancelled if the context deems it appropriate. Let us consider the examples below:

(3) a. She ate some of the grapes.
b. She ate some of the grapes. In fact, she may have eaten all of them.

The sentence in (3-a) contains the scalar term *some*. In Chierchia's (2004a) grammatical view of scalar implicature, the speaker utters (3-a), which contains the default semantic value *some but not all*. Then, the hearer accurately hypothesizes that the strengthened interpretation was intended. When the hearer encounters an utterance such as (3-b), they process the first sentence with the strengthened semantic meaning, but then the hearer processes the second sentence and rejects the strengthened interpretation. This approach to scalar implicature is based within the grammatical system as opposed to the pragmatic theories of scalar implicature that rely on "extragrammatical modules" (Chierchia, 2004a, p. 70).

In all of the theories described above, regardless of their different methods for outlining the pragmatic reasoning process behind scalar implicature, there is an appeal to alternative expressions. Which implicatures arise crucially depends on which alternatives a listener believes to be activated. All of these theories consider relevance when determining whether an alternative is activated. Which is to say, only semantically relevant alternatives are activated when calculating a scalar implicature. These theories differ, however, in the additional constraints they impose on the set of activated alternatives. Some theories say complexity constrains the set of alternatives, where only alternatives that are less than or equal in complexity are active and available to interlocutors. Other theories consider prevalence when determining which alternatives are activated – alternatives that are more colloquial are available to listeners. In the next subsection, I outline these different approaches to constraining the set of alternatives.

2.2 Constraints on the set of alternatives

In addition to breaking down the pragmatic reasoning process behind implicatures, multiple theorists have proposed constraints for determining what utterances can be considered in the set of alternatives. Because the reasoning process(es) outlined above rely on the existence of a stronger alternative, determining what utterances exist as alternatives along the same scale is a crucial part of understanding which scalar implicatures are calculated. The lack of the scalar implicature *some* $\rightsquigarrow$ 'not X' can be explained by the assumption that X is not an available alternative to speakers. Most theories of scalar implicature consider the complexity of an utterance as a predictor of which relevant utterances are available as alternatives – more complex utterances are typically ruled out. Some theories of scalar implicature also consider the prevalence of an utterance when determining whether or not it is present in the set of available alternatives (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Alternatives that are more colloquial, i.e. have higher prevalence, are included, while less prevalent alternatives are excluded.

Under the Cooperative Principle, Grice (1975) includes the maxim of manner, which leads to the idea that complexity plays a role in scalar implicature. Under the maxim of manner, Grice also includes the submaxim "be brief" (p. 46). This rule provides general guidelines that seem to suggest that complexity is a contributing factor to scalar implicature calculation. Horn (1984, p. 34) elucidates what Grice (1975) says, stating that utterances narrow in opposition to other utterances on the same scale via the maxims of Quality and Manner.

Therefore, given an utterance such as *She ate some of the grapes*, the hearer thinks: if the speaker meant 'all' they would say *all*, which is just as complex, obeying the maxim of manner, and more informative, obeying the maxim of quantity. Horn's (2000) approach to scalar implicature imposes restrictions on the set of alternatives based on complexity or lexicalization. Alternative utterances are only activated if they are equal or lesser in complexity *or* if they are a lexically specified alternative (2000).

Katzir (2007) presents a framework for computing scalar implicature that calculates the set of alternative utterances based on a definition of syntactic complexity. In fact, Katzir (2007) argues that structural complexity is the only relevant factor for determining what restricts the set of alternatives (p. 688). This strong view of the role of complexity was developed by Katzir in response to the Symmetry Problem (von Fintel & Fox, 2002). The use of *some* implicates the negation of the stronger, unused alternative *all*. There exists, however, another even stronger alternative *some but not all* that the speaker also did not produce. If this alternative utterance were included in the set that is available to interlocutors, we would expect the inaccurate implicature *some* implicating the negation of *some but not all*. Therefore, the symmetry problem is the problem with including *some* in the set of alternatives while also excluding *some but not all* (Breheny et al., 2018). According to Katzir (2007), whether or not an alternative is included is categorical in nature: if an utterance is "at most as complex" as ϕ *or* it is an item in the lexicon, it is present in the set of alternatives, $A\phi$. If an utterance is more complex than ϕ or it is *not* present in the lexicon, it is ruled out as an alternative. The alternative *some but not all* would therefore be excluded from the set of alternatives because it is more complex than *some*.

Swanson (2010) argues against a role of complexity altogether, stating that using complexity to restrict alternatives, as proposed by Katzir (2007), is more restrictive than necessary. The examples below are taken directly from Swanson (2010). These examples present a series of alternatives that do not differ in their structural complexity.

(4) a. The heater sometimes squeaks.
b. $\rightsquigarrow$ The heater intermittently squeaks.
c. $\rightsquigarrow$ The heater occasionally squeaks.
d. $\not\rightsquigarrow$ The heater constantly squeaks.

According to Katzir's account, if a speaker chooses to say (4-a), we would predict that the listener will conclude that the speaker does not believe the more informative utterances in (4-b) and (4-c). This line of reasoning would calculate (4-d) as an implicature: since *intermittently* and *occasionally* are available alternatives, the listener will inaccurately conclude that the speaker believes they are false since they were not uttered. Swanson (2010) argues that complexity fails to account for why *sometimes* is used when *occasionally* is an available alternative for the speaker, since *occasionally* is not ruled out based on complexity. It is important to note that while Swanson (2010) argues directly against Katzir's (2007) approach to constraining the set of alternatives, he does not propose an alternative argument for calculating alternatives that does not give rise to the symmetry problem.

Unlike Horn (2000) and Katzir (2007), whose theories constrain the set of alternatives from a complexity perspective, Geurts (2011) imposes constraints based on complexity and prevalence. Geurts (2011) explains that under the Standard Recipe there are "substantial" constraints on the available alternatives. Namely, if an alternative is more complex, i.e. longer, it will not be available to the speaker. He takes this one step further by considering how colloquial an alternative is. According to Geurts (2011), an alternative with higher prevalence – which he refers to as “availability” – is contained within the set of alternatives, while an alternative with low prevalence is excluded (p. 121). He presents the example of *dog* in English.

(5) Lee-Ann took the dog to the park.

When a speaker utters a sentence such as (5), the use of *dog* does not implicate that the speaker is unaware of the specific breed of dog. This is in stark contrast to the use of other weaker forms like *some*. Geurts (2011) argues that specific dog breeds are less available to English speakers than *all*, for example, so they are not activated when someone utters the weaker form *dog*.

The Rational Speech Act (RSA) framework considers both complexity and prevalence when outlining the implicature calculation process (Goodman & Stuhlmüller, 2013). In contrast to the theory outlined by Katzir (2007) and Geurts (2011), the RSA framework suggests a gradient approach to determining whether or not a given alternative is included. This is a different approach to constraining the set of alternatives since complexity and prevalence still play a role, but they do not rule out alternatives entirely for not adhering to the rules of complexity outlined by Katzir (2007) for example. The model assigns probabilities to alternative utterances that align with how prevalent an utterance is – the more colloquial, or common, an utterance is, the higher prior probability it has. The framework also considers complexity when determining these probabilities. A cost parameter (β) is applied to each alternative utterance, with more complex utterances receiving higher “costs”. The result is a set of relevant alternatives that each have a probability between 0 and 1 – more costly utterances (i.e. alternatives with more words) have lower probabilities than less costly utterances, but utterances with higher prevalence have higher probabilities. Prevalence and complexity work together within the RSA framework.

Each of these theories of scalar implicature propose different methods for constraining the set of alternatives. Horn (2000) and Katzir (2007) restrict the set of alternatives based on complexity. Geurts (2011) incorporates complexity into his theory of scalar implicature, but he also includes prevalence as a constraint on which alternatives are activated. The RSA framework combines complexity and prevalence to rank the set of relevant alternatives available to interlocutors. Nevertheless, there has been little cross-linguistic research empirically investigating the role of prevalence in restricting alternatives and calculating scalar implicatures. The following section details a specific hypothesis Horn (2000) makes about the scalar implicature *finger* ⇝ 'not thumb', which is the foundation for my study.

2.3 Digits in Scalar Implicature

To make things more concrete let us concentrate on a specific example: *finger* ⇝ 'not thumb'. Horn (2000) highlights an asymmetry that exists between *thumb* & *finger*, and *big toe* & *toe*. The thumb is generally considered a type of finger, but it seems that uttering *finger* conveys 'not thumb', since *thumb* is an alternative to *finger* (6-a). Using the term *toe*, however, does not implicate 'not big toe' (6-b). Horn (2000) explains that *thumb* crucially exists as a "viable" alternative to *finger* in a way that *big toe* does not exist for *toe* (p. 308).

(6) a. I hurt my finger. ⇝ I did not hurt my thumb.
    b. I hurt my toe. ⇝̸ I did not hurt my big toe.

Horn (2000, p. 308) continues: "We would predict that if the colloquial language replaced its *thumb* with the polymorphous *pollex* (the Latin and scientific English term for both 'thumb' and 'big toe'), the asymmetry [between *finger* and *toe*] would instantly vanish". Based on Horn's line of pragmatic reasoning, there should be no narrowing of *toe* in opposition to *big toe*. Since *big toe* is more complex than *toe*, Horn would rule it out as a possible alternative.

In a follow-up to Horn's *pollex* prediction, Geurts (2011) zeroes in on Horn's strategic use of the term "colloquial" when he says "It is important to note, however, that the adjective 'colloquial' is doing real work in this statement: it is not enough for an alternative word to be in the language; it has to be sufficiently salient, as well: if the word 'thumb' was rarely used, then presumably the asymmetry between [finger and toe] would vanish too" (p. 122). That is, if a stronger utterance is present in the language *and* more salient to speakers than another strong utterance, a scalar implicature will arise if the weaker form is used. Geurts (2011) agrees that the concept of "viability" plays a role in scalar implicature calculation, and he argues that "viability" is determined based on more than just complexity – namely, whether or not an alternative is "colloquial".

Horn (1984) has discussed the robust nature of certain implicatures, positing a diachronic change, called "Q-based narrowing". Before complete Q-based narrowing can take place, the more general term becomes a hyponym of itself ('autohyponym'). At this intermediate step, the general meaning of the term is still preserved in some contexts, but an implicature typically arises when it is used in conversation. English *rectangle* and *finger* are thought to be in this intermediate stage. Thus, it is not entirely clear whether it is the scalar implicature *finger* $\rightsquigarrow$ 'not thumb' or the term *finger* that has a narrow semantic meaning that excludes the 'thumb'. Although Horn claims that *finger* is autohyponymous, his prediction seems to set aside the narrower *finger*. The prediction that the asymmetry disappears only follows if *finger* only has the broad sense (and *toe* does too – which was not in question). Nevertheless, the present study sets out to test a prediction that Horn makes which relies upon the assumption that the narrower term *finger* is not being used.

As shown in the previous sections, the exact role of prevalence is still a generally wide open area for research. Something that has not been investigated systematically, but which is a natural place to begin research, is in cross-linguistic comparative pragmatics because different languages differ in the range of alternatives they make available. More specifically, Horn (2000) and Geurts (2011) have made specific claims that can be empirically investigated through cross-linguistic research. I can study the properties of the semantically associated items in the lexicon as a function of the prevalence of the alternatives within the same semantic domain. As Geurts (2011) states, if prevalence of alternatives does play a role, then one would predict cross-linguistic differences in scalar implicature computation. In languages where the translational equivalents for *thumb* are less prevalent, it is predicted that an implicature from the more general term *finger* to ‘not thumb’ is less likely to arise. The present study tests this prediction with Spanish, where the translational equivalent of *thumb* is less prevalent (i.e. less colloquial). Spanish contains widespread variation and does not have a conventionalized single-word translational equivalent for *thumb*, or any of the other fingers. In fact, there are multiple phrases that exist in Spanish to convey the meaning *thumb*: *pulgar* ‘thumb’, *dedo gordo* ‘fat digit’, *dedo pulgar* ‘thumb digit’ to name a few.

Horn’s (2000) aside coupled with Geurts’s (2011) follow-up provide the foundation for the research questions (see Section 4) and the present study: Spanish contains a single-word alternative *pulgar* – the Spanish descendant of Latin *pollex* – that is believed to be significantly less prevalent than *thumb* is in English. Crucially, *pulgar* does not differ from *thumb* in complexity, but it does differ in prevalence. *Pulgar* ‘thumb’ is a hyponym to *dedo* ‘finger’ without the hand/foot distinction we see with *thumb* and *finger*. This slight difference in definition is not important for the present paper; the crucial feature of *pulgar* that is under investigation is that it is a less “viable” alternative for Spanish speakers. If the theory proposed by Geurts (2011) is correct, the asymmetry between *finger* and *toe* that is believed to exist in English is predicted to be absent in Spanish.

The four experiments in the present study test these predictions. I conducted production experiments each in English and Spanish as norming studies to gain insight into the prevalence of available alternatives. The production experiment methodology and results are described in Section 3 below. I then outline my research questions in Section 4. Section 5 describes the relevant predictions for the comprehension studies, and Section 6 outlines the comprehension studies and their results.

3 Production Studies

I conducted two production studies: one in English and one in Spanish. As mentioned above, the production studies act as norming studies to measure the prevalence of available alternatives. Both tasks contained images of body parts with tattoos (see Figure 1).

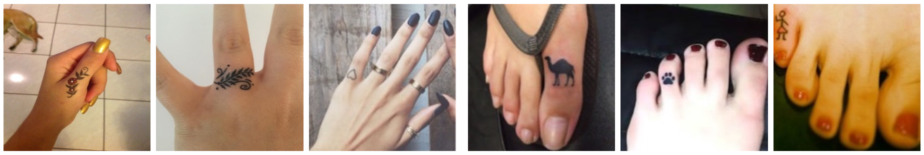


Figure 1: **Stimulus Items for Production and Comprehension tasks**

3.1 Methodology

3.1.1 Participants

All participants were recruited on Prolific, an online recruitment platform for web-based experiments. All studies involved different groups of participants. For each study, participants were filtered for a number of demographic features: native language, country of birth, current country of residence, and whether they were raised monolingual. In the English production study, all participants were self-reported monolingual native English speakers who were born and currently live in the United States. In the Spanish production study, all participants were self-reported monolingual native Spanish speakers who were born and currently live in Mexico. There were 24 American English speakers and 23 Mexican Spanish speakers in the production studies.

3.1.2 Materials

Participants completed a task in which they were asked to look at a series of pictures. All of the pictures were body parts with a tattoo on them. The tattoos served as an indicator of which digit or body part the speaker was talking about. The target items showed photos of a tattoo on the thumb, ring finger, pinky, big toe, fourth toe and pinky toe.

Thumb was chosen based on the scalar implicature discussed by Horn (2000) and Geurts (2011) discussed above. *Ring finger* was chosen because it is more complex than *thumb*. *Pinky* was chosen because it is a single-word term for another finger, but does not seem to be excluded from the set of possible referents when someone says *She has a tattoo on her finger*. The toe counterparts to all three target finger terms were chosen to investigate whether the scalar implicature *toe* $\rightsquigarrow$ 'not big toe' exists.

I included six filler items: two photos of tattoos on the leg, two photos of tattoos on the arm, and two photos of tattoos on the back. Although there were only three body part categories for the filler items, each of the filler items depicted tattoos in different locations. For example, one arm filler item depicted a tattoo on the shoulder while the other arm filler item depicted a tattoo on the elbow.

3.1.3 Procedure

Participants were shown a series of photos, one by one. Once they were given the photo, they were then asked to fill in the blank of the sentence: *She has a tattoo on* ______, or its translational equivalent, in the case of Spanish. The order of images was randomized. All participants were presented with all six target items and all six filler items.

The production studies acted as a norming study to determine what alternatives speakers use when referring to the target digits. These production results will act as a measure of prevalence of the alternatives available to speakers given the general terms *finger* and *toe.*

3.1.4 Normalizing production results

After the data collection process was completed, all responses for the production study were normalized. This included removing additional words such as "left" or "right" (e.g. "right pinky" became "pinky"). Directional terms, articles, and other non-essential words were stripped away so that all that was remaining was the word or phrase that was used to refer to the digit itself. This removed excess noise from the data and allowed me to group responses together that were essentially identical in form – *dedo de la mano* ('digit of the hand' - *finger*) vs. *dedo* ('digit'), for example. Additionally, responses were coded for specificity — 1 for specific words/phrases that could refer to only one digit (e.g. "thumb" or "pulgar") and 0 for non-specific words/phrases that could refer to more than one digit (e.g. "finger" or "dedo de la mano"). Responses were also coded for directionality, that is to say if a participant used "left" or "right" in their response. That data is outside of the scope of this paper, and thus will not be discussed in further detail. The results for the production studies are presented below.

3.2 Results

In the production study, when participants were shown the image with a tattoo on the thumb and were asked to fill in the blank: "She has a tattoo on her ______.", 100% of English speakers responded with *thumb* — a specific term. In contrast, Spanish speakers were not unanimous in their responses, and in fact exhibited much more variability. Table 1 below presents the production results for all digits on the hand. While the single-word translational equivalent to 'thumb', *pulgar*, was preferred over all other utterances, only approximately 42% of participants used it. *Mano* – Spanish for 'hand' – was the second most frequent response with 17.4%. Spanish speakers preferred using specific terms 63.2% of the time, while 34.8% of the participants used a general term.

For the ring finger item (presented in Table 1), English speakers preferred the specific term *ring finger* 83% of the time, but some participants (17%) did produce the general term *finger*. Again, Spanish speakers presented more variation in their responses than English participants, with seven unique utterances produced. 35% of Spanish participants produced *dedo anular* ('ring finger'); 26% produced *dedo* ('finger'). Overall, Spanish participants trended like English participants with preference for specific (52.2%) over general terms (47.8%), although the preference was not as strong (see Table 1).

English production for the pinky acts similarly to English production for the ring finger (see Table 1) – there is variation between specific and general

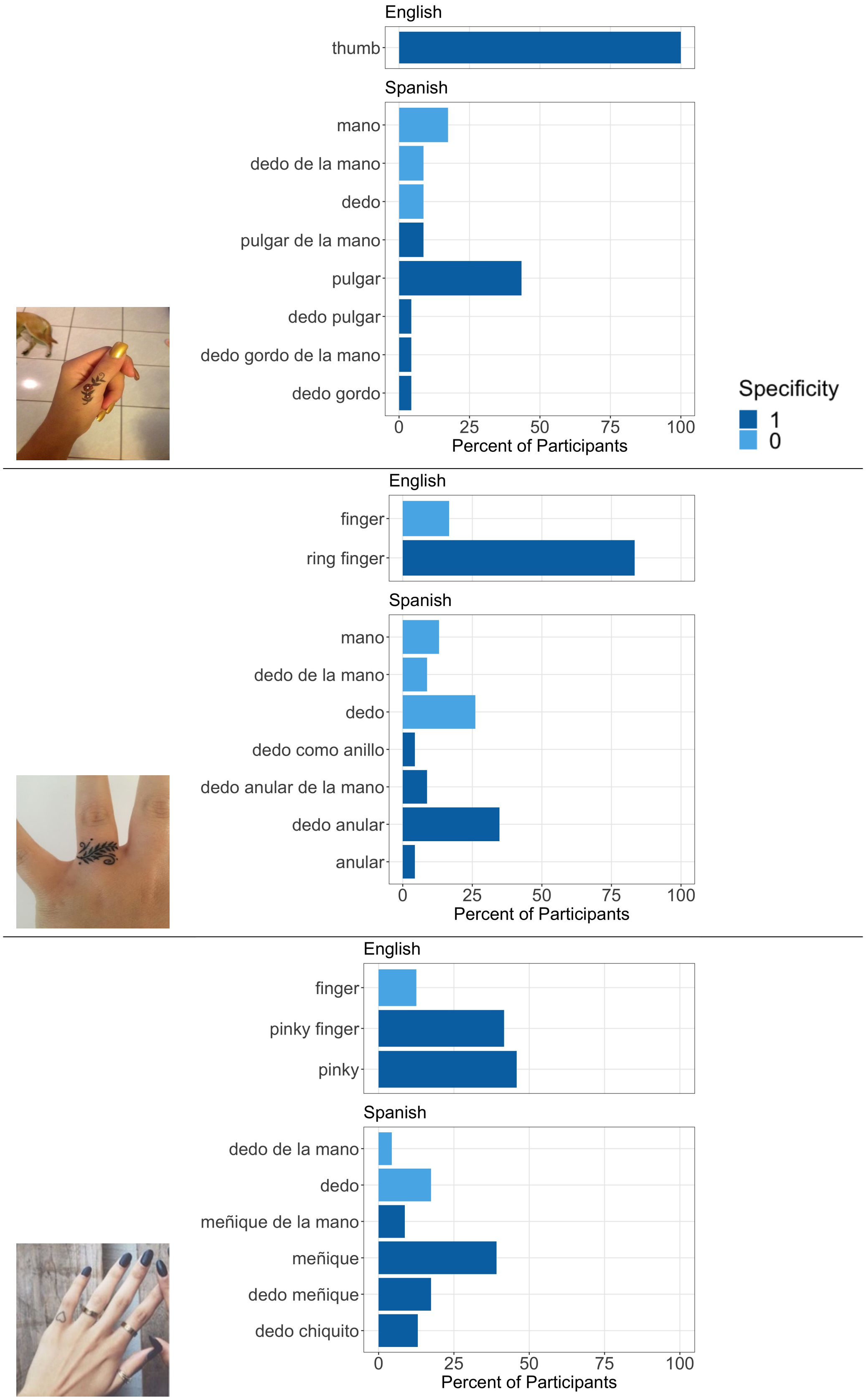


Table 1: English and Spanish production of finger terms;“specific” terms refer to a single digit

term usage, with 12.5% of participants preferring the general term *finger* over a specific form (e.g. *pinky* – 45.8%, or *pinky finger* – 41.7%). The single word *pinky* is used (and most frequently), yet almost as many participants chose to use the two-word alternative *pinky finger*. Spanish speakers also present varying responses. The most common utterance was the single-word equivalent for ‘pinky’, *meñique*, at 39.1%. 21.7% of Spanish speakers utilized a general form over a specific one, which shows a higher preference than was present in our English population.

For the big toe item (presented below), English speakers favored the specific term *big toe* 83.3% of the time over the general term *toe*, which was only used 16.7% of the time. Unlike English speakers, who only used two different terms to refer to the location of the tattoo, the Spanish participants produced six different descriptions (shown below in Table 2). The most common response (69.6%) was the specific term *dedo gordo del pie* (lit. ‘fat digit of the foot’), or ‘big toe’. The second most common response at 34.8% was the general term *dedo del pie* (lit. ‘digit of the foot’), which translates to *toe*.

When participants were presented with an image of a tattoo on the ring toe, English participants showed increased dispersion in their responses. The majority of participants (58.3%) used the general term “toe”. The remaining participants (41.7%) produced various specific terms for the digit (“fourth toe” and “ring toe” to name a few). Spanish speakers had a much higher rate of general term usage, with 82.6% of participants preferring terms like *dedo del pie* ‘toe’, *dedo* ‘digit’, or *pie* ‘foot’.

Finally, the production results for the pinky toe present similar trends in English and Spanish. English speakers preferred using the general term *toe* far less than a specific term (20.8% and 79.2%, respectively). There was less dispersion in English production results for the pinky toe than for the ring

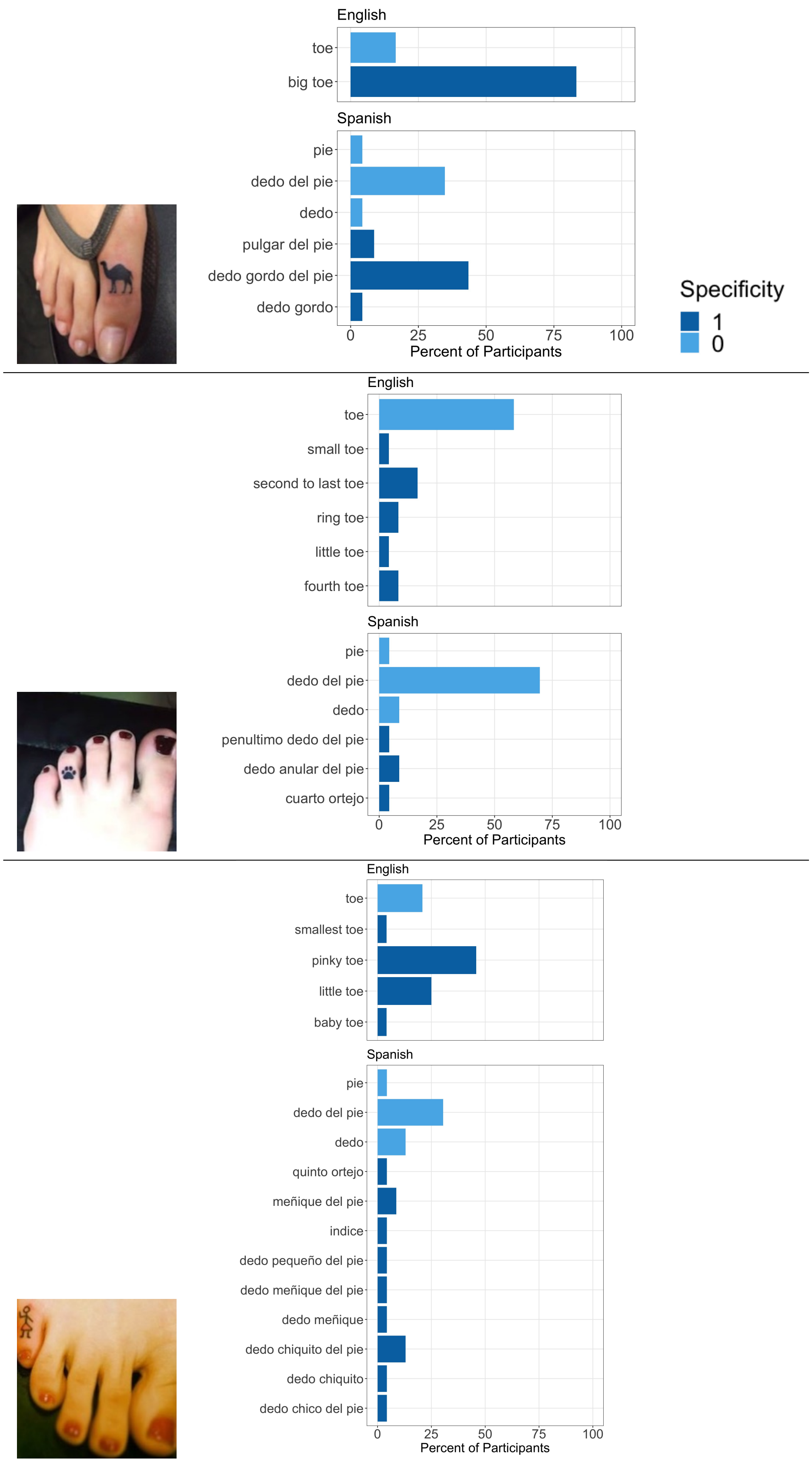


Table 2: English and Spanish production of toe terms; "specific" terms refer to a single digit

toe. In Spanish, participants generally preferred a specific term (52.2%) over a general term (47.8%), but the trend was not as strong as in English. In contrast to English, Spanish production data presented a much larger amount of dispersion for the pinky toe than the ring toe, as shown in Table 2.

4 Research questions and Predictions

4.1 Research Question

The cross-linguistic differences we see in speakers' production show that alternative utterances for each digit differ in how prevalent they are in English compared to Spanish. These production results support Horn's (2000) speculation – the Spanish single-world alternative *pulgar* 'thumb' is less prevalent than *thumb* is in English. This suggests that Spanish speakers and English speakers should differ in scalar implicature calculation due to the differences in prevalence – and complexity — of the alternative forms for 'thumb'. Thus, my research question is as follows: When asked to choose between two digits as referents for a general term, do English and Spanish speakers prefer one digit over the other in accordance with the prevalence associated with the specific terms for that digit, or with the complexity associated with the specific terms for that digit?

The comprehension studies, outlined in Section 5, provide insight into what scalar implicatures English and Spanish speakers calculate with regard to the digits. Complexity-based theories for constraining the set of alternatives predict different results than prevalence-based theories for constraining the set of alternatives. I outline these specific predictions in the next subsection.

4.2 Predictions

The predictions for the present comprehension studies follow the hypotheses of Horn (1984) and Geurts (2011). A prevalence-based constraint on alternatives would predict that English and Spanish will differ with respect to the scalar implicature associated with*finger*. Prevalence-based theories predict that unlike English speakers, Spanish speakers will not show evidence for the scalar implicature *finger* $\rightsquigarrow$ 'not thumb' when presented with an image pair with a thumb and ring finger. This is due to the lower prevalence of the term *pulgar* 'thumb', which is supported by the production results. However, complexity-based theories would predict an implicature in both languages, since the translational equivalent for 'thumb' in Spanish is equal in complexity. Thus, if the comprehension results reveal that Spanish speakers *do* calculate the implicature *finger* $\rightsquigarrow$ 'not thumb', this would support theories that constrain the set of alternatives due to complexity.

If speakers are constraining the set of alternatives based on prevalence, it is predicted that speakers in English and Spanish will not show evidence of the scalar implicature *finger* $\rightsquigarrow$ 'not pinky' given the image pair containing a pinky and a ring finger. This is because the production results reveal that *pinky* is not as prevalent an alternative in either language. In the English production results, *pinky* competes with the more complex alternative *pinky finger*, and in Spanish, there are a large number of alternatives that are in competition with *meñique* 'pinky'. Complexity based theories would predict that English and Spanish speakers will calculate the implicature *finger* $\rightsquigarrow$ 'not pinky', since both languages have a single word alternative for 'pinky'. Thus, If English and Spanish speakers do calculate this implicature, this would support theories that constrain the set of alternatives based on complexity.

Additionally, prevalence-based theories support the prediction that participants will not calculate a scalar implicature for any of the digits on the feet in both English and Spanish. The production results suggest that none of the specific terms are prevalent enough to block the general term *toe* in English and *dedo del pie* 'toe' in Spanish. This prediction follows directly from the asymmetry that Horn (2000) outlines. Complexity theories would predict speakers will not calculate the implicature *toe* $\rightsquigarrow$ 'not pinky toe', since *pinky toe* is more complex than *toe*. This prediction carries for the other specific terms for the toes (*big toe* and *ring toe*) too.

In sum, if participants in the comprehension tasks preform in a way that mirrors the production results, I can conclude in favor of theories that constrain the set of alternatives based on prevalence. That is to say that prevalence-based theories would hold if participants calculate implicatures for digits that presented higher rates of general term usage in the production tasks (i.e. digits where the specific terms are less prevalent). However, if participants calculate implicatures based on the production-based predictions described above, I can conclude against theories that impose constraints on alternatives based on prevalence.

5 Comprehension Studies

The comprehension tasks provide a measure of the extent to which speakers of each language are computing scalar implicatures associated with the general terms for 'finger' and/or 'toe' in English and Spanish.

5.1 Methodology

5.1.1 Participants

45 American English participants and 48 Mexican Spanish participants completed the comprehension task. They were also recruited from Prolific. Like the participants from the production tasks, participants were pre-screened for language, monolingual status, country of birth and country of residence. English participants were self-reporting American monolinguals that were born and currently reside in the United States. Spanish participants were self-reporting Mexican monolinguals that were born and currently reside in Mexico.

5.1.2 Materials

The target items for the comprehension studies consisted of 6 image pairs. Three of the pairs were images of hands and three of the pairs were images of feet such that all possible hand combinations and all possible foot combinations were presented. No target pairs consisted of an image of a digit on the hand and an image of a digit on the foot. The images were the same six images from the production study (see Figure 1).

In addition to the 6 target image pairs, participants were also presented with 6 filler image pairs. Three of the filler pairs were "easy", where participants were given an utterance (e.g. *She has a tattoo on her back*) and were presented with a pair of images that contained only one back tattoo. The other three filler pairs were considered "hard" because these image pairs contained two different back tattoos, for example. Filler pairs that were "easy" acted as attention checks, since there was a clear correct response. Participants that failed one or more "easy" fillers were eliminated from the results. To eliminate an effect of survey versions, item order and left-right presentation of the images were randomized in the study.

5.1.3 Procedure

Participants were asked to choose from a pair of images based on an utterance of the form *She has a tattoo on her finger/toe/dedo ('finger')/dedo del pie ('toe')/etc.* The referents were presented as image pairs: thumb on the left, ring finger on the right, for example. These tasks were forced-choice, asking participants "Which picture are they talking about?" and requiring them to click on the image on the left or the image on the right to advance to the next item.

5.2 Results

For the comprehension study, responses were simply coded as the image the participant clicked on (e.g. "thumb" for the image with the tattoo on the thumb). The p-values I report are the result of conducting a 1-sample proportion test, where the null hypothesis, or the probability of choosing the correct image is 0.5. The assumption is that the data follow a Bernoulli distribution. I then ran a BH adjustment on the p-values. The comprehension results provide insight into whether speakers calculate a scalar implicature when presented with a pair of images.

In the comprehension study, when participants were asked to choose between the thumb image and the ring finger image given the statement "She has a tattoo on her finger", 75% of English participants chose the image of the ring finger, $p = 0.004$ (see Figure 2). In contrast, just over half of the Spanish speakers chose the ring finger image over the thumb image, but the error bar, which depicts a 95% Confidence Interval, distinctly crosses the 50% mark, showing that the Spanish participants' responses are not statistically significantly different from chance ($p = 0.627$).

In contrast, when English participants were asked to choose between the

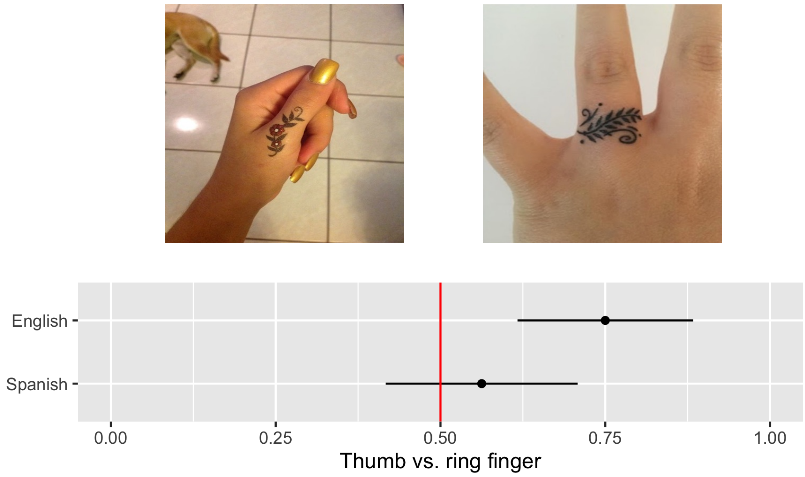


Figure 2: Observed frequency with 95% CI of choosing Thumb or Ring Finger in English and Spanish

ring finger image and the pinky finger image given the utterance “She has a tattoo on her finger”, slightly more than half of the participants chose the ring finger over the pinky finger. This result was not significantly different from chance, p = 0.781 (see Figure 3). Spanish participants showed a slight trend toward the pinky finger, but this result was not significantly different from chance (p = 0.965).

As shown in Figure 4, when tasked with choosing between the thumb and pinky finger, over 75% of English participants chose the pinky image given the utterance “She has a tattoo on her finger” ($p < 0.001$). While Spanish participants showed a slight preference for the pinky finger image over the thumb image, the error bar does cross the 50 % mark, suggesting that the tendency is not significantly different from chance (p = 0.145).

For the big toe and ring toe image pair, English participants showed a slight preference for the ring toe image, with roughly 63% of participants choosing that image. However, the error bar indicates that this result is not statistically different from chance (p = 0.145). Spanish participants actually showed

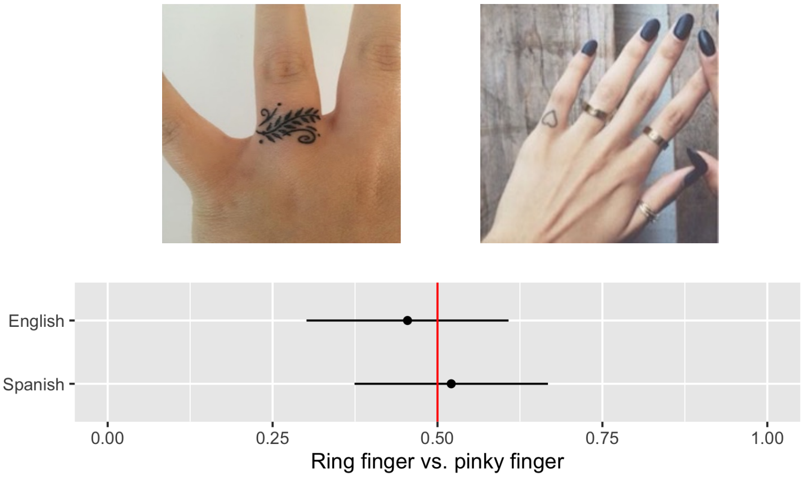


Figure 3: Observed frequency with 95% CI of choosing Ring Finger or Pinky Finger in English and Spanish

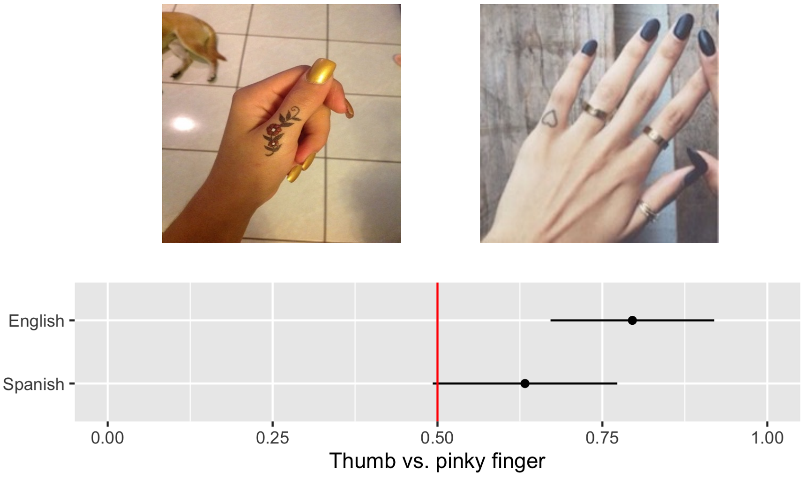


Figure 4: Observed frequency with 95% CI of choosing Thumb or Pinky Finger in English and Spanish

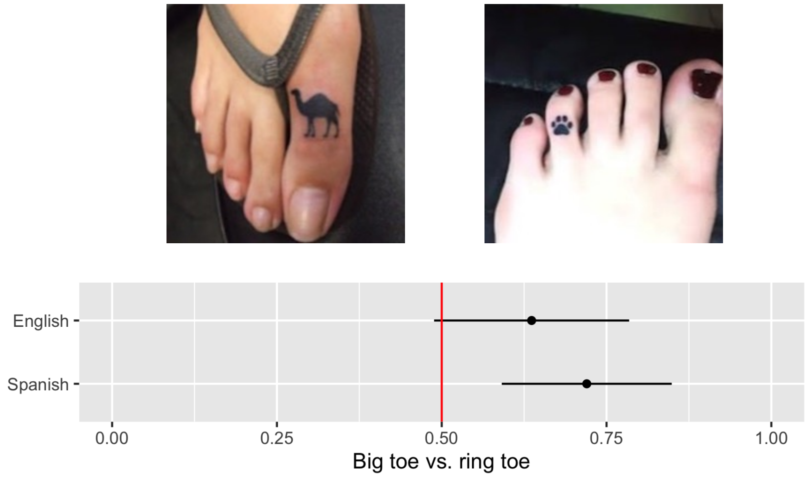


Figure 5: Observed frequency with 95% CI of choosing Big Toe or Ring Toe in English and Spanish

a stronger trend toward the ring toe given the translational equivalent of “She has a tattoo on her toe” ($p = 0.007$). This is shown in Figure 5 below.

When English participants were presented with the big toe and pinky toe image pair, participants preferred the big toe image given the utterance “She has a tattoo on her toe”, but the difference was not statistically significant, $p = 0.10$ (see Figure 6). Spanish participants, however, were completely divided in their responses. 50% of participants chose the big toe image, showing that Spanish speakers do not calculate an implicature ($p = 1.0$).

The comprehension results for the ring toe and pinky toe pair revealed similar results for English participants (p = 0.002) and Spanish participants (p = 0.0009). When speakers in each language were given the utterance “she has a tattoo on her toe”, roughly 75% of participants believed the speaker was referring to the ring toe as opposed to the pinky toe.

In summary, the comprehension results suggest that implicatures were computed in the context of English *thumb* vs. *ring finger*, *thumb* vs. *pinky finger*, and *ring toe* vs. *pinky toe*. The results also suggest that Spanish speakers cal-

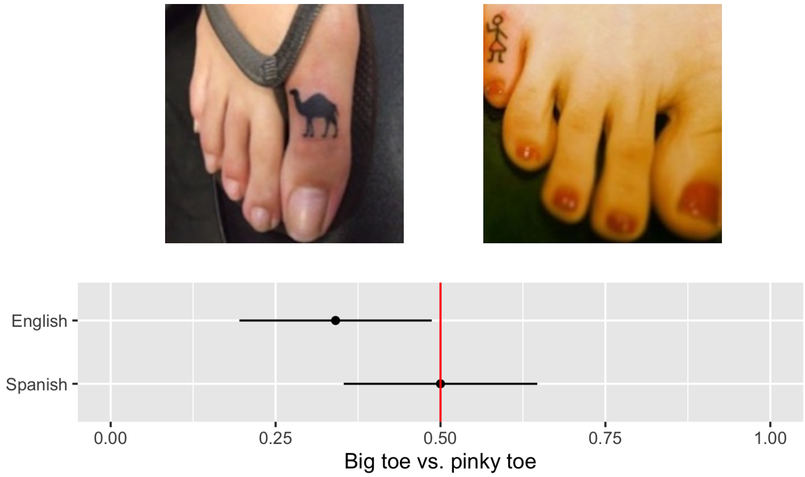


Figure 6: Observed frequency with 95% CI of choosing Big Toe or Pinky Toe in English and Spanish

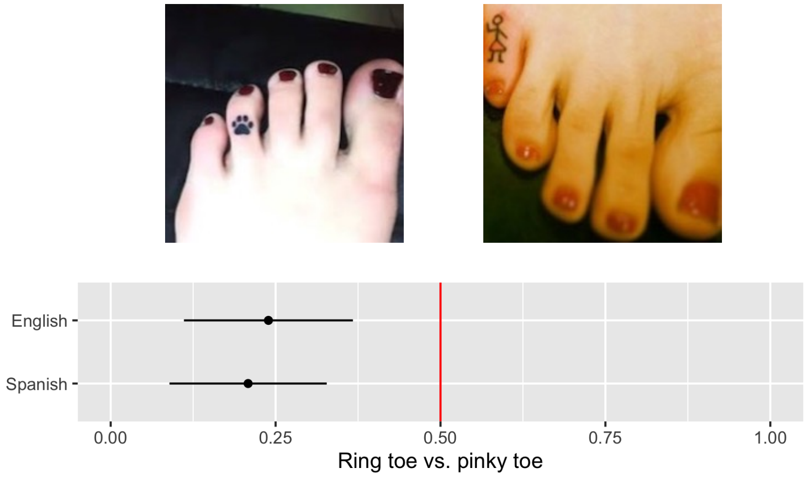


Figure 7: Observed frequency with 95% CI of choosing Ring Toe or Pinky Toe in English and Spanish

culated scalar implicatures in the context of *big toe* vs. *ring toe* and *ring toe* vs. *pinky toe.* These results are reflected in Table 3 below. The estimate is the estimated true proportion in the greater population. I conducted a Benjamini-Hochberg adjustment to obtain the adjusted p-values.

	Condition	Language	Estimate	p-value	adj. p-value
1	Big toe vs. ring toe	Eng	0.64	0.097	0.145
2	Big toe vs. ring toe	Spa	0.72	0.002	0.007*
3	Big toe vs. pinky toe	Eng	0.34	0.05	0.10
4	Big toe vs. pinky toe	Spa	0.50	1.00	1.00
5	Ring toe vs. pinky toe	Eng	0.24	0.0006	0.002*
6	Ring toe vs. pinky toe	Spa	0.21	0.00009	0.0009*
7	Thumb vs. ring finger	Eng	0.75	0.002	0.004*
8	Thumb vs. ring finger	Spa	0.56	0.47	0.627
9	Thumb vs. pinky finger	Eng	0.80	0.0001	0.0009*
10	Thumb vs. pinky finger	Spa	0.63	0.086	0.145
11	Ring finger vs. pinky finger	Eng	0.45	0.651	0.781
12	Ring finger vs. pinky finger	Spa	0.52	0.885	0.965

Table 3: p-values and adjusted p-values for each language/condition pair.

6 Bayesian Pragmatics

To gain a better understanding of complexity and prevalence, and their roles in scalar implicature, I compared two different speaker models within the context of RSA. The first model is a more traditional speaker model that penalizes longer – more complex – utterances (now referred to as the Complexity model). The second model is a prevalence-based speaker model that has perfect knowledge of speaker production (now referred to as the Production model). For both models, I incorporated six underlying states that correspond to the six target digits (*thumb, ring finger, pinky, big toe, ring toe,* and *pinky toe*). Literal meanings for each utterance from the production study were hand-specified as a set of states for English and Spanish (see Appendix).

For the complexity-based speaker model, as presented earlier, the Speaker chooses an utterance based on accuracy and cost. Length is equivalent to length in words, and $L_0(s\,|\,u)$, L_0 of u given s, is the probability that a literal listener will choose a state s given an utterance u. The model contains two free parameters. Alpha (α) is known as the rationality parameter, which corresponds to how accurate a speaker would like to be when communicating to a listener (how much the model cares about accuracy). The second parameter beta (β) is the cost, which is multiplied by the number of words in the utterance. The cost parameter corresponds to speakers' preference to be as concise as possible when speaking. Model parameters for the Complexity model were tuned to the thumb/ring finger data point from the experimental results of the English comprehension study: α set at 1 and β set at 2.

$$S(u\,|\,s) \propto \mathsf{exp}(\alpha \cdot L_0(s\,|\,u) - \beta \cdot \mathsf{length}(u))$$

A pragmatic listener was then built on top of the Complexity speaker model. As described in section 2.1, the pragmatic listener chooses an interpretation based on the speaker:

$$L(s\,|\,u) \propto S(u\,|\,s) \cdot P(s)$$

In contrast to the Complexity model, the prevalence-based speaker model is fed the exact production probabilities for each utterance collected from the production study. This ensures that the Production model has full awareness of what utterances are more or less prevalent for speakers – these are utterances speakers actually produced.

$$S(u\,|\,s) \propto F(u\,|\,s)$$

The prevalence-based speaker chooses an utterance based on the empirically observed frequencies in my production data, where $F(u\,|\,s)$ is the frequency with which an utterance u was used in the production experiments to describe state s (i.e. the finger or toe that had the tattoo). Like the Complexity model, a pragmatic listener is coded on top of the Production model. Thus, when the Pragmatic Listener is considering the Speaker, S, it has full awareness of speaker production. The Pragmatic Listener model is the same for both Speaker models.

6.1 Model performance

The model predictions in comparison to the empirical results are presented in Figure 8. The Complexity model inaccurately predicts no implicature for the thumb/pinky item in English. This is because the model is only considering the fact that these two utterances are equally complex. The model prediction for the thumb/pinky item in Spanish falls within the margin of error for the empirical results. Additionally, the model incorrectly predicts that there will be an implicature for ring finger/pinky in English since the one-word term *pinky* is an available alternative. The model accurately predicts no implicature in Spanish for the ring finger/pinky item. Because the Complexity model was tuned to the English thumb/ring finger item, the model accurately predicts an implicature in English. It also accurately predicts no implicature in Spanish.

For digits on the feet, the model incorrectly predicts no implicature for ring toe/pinky toe in English and big toe/ring toe in Spanish – since they are equally complex. However, the production results suggest that they are not equally viable as alternatives. Since *ring toe* and *pinky toe* are equally as complex, if complexity alone determined which alternatives were available to speakers, we would expect no implicature to arise. However, the presence of the implicature *toe* $\rightsquigarrow$ 'not pinky toe' suggests that *pinky toe* is a more prevalent alternative than *ring toe*. These results suggest that something else is going

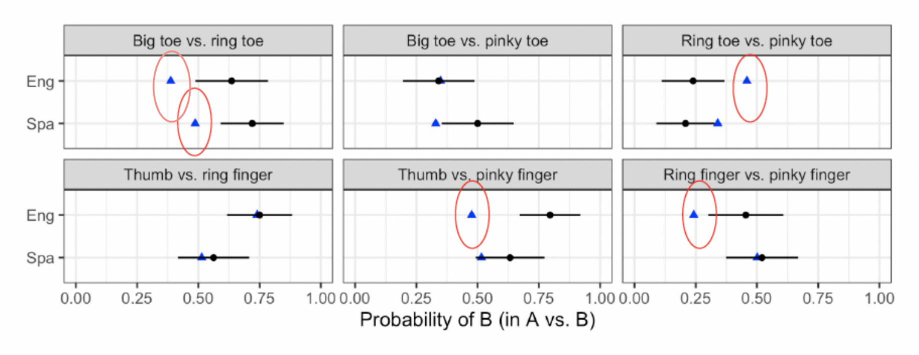


Figure 8: Complexity Model predictions (triangle) plotted against comprehension results; Inaccurate model predictions are circled in red

on in calculating implicatures that complexity alone cannot account for. The complexity-based model fails to understand that two alternatives with equal complexity may differ in their prevalence. The Production model, on the other hand, is capable of accounting for this.

Overall, the Production model seems to perform much better (see Figure 9) than the Complexity model. It only incorrectly predicts no implicature for big toe/pinky toe in English. The Production model predicts stronger implicatures for the thumb/ring finger items in English and Spanish, and for the thumb/pinky finger items in English. Which is to say, where the comprehension results trend rightward, at just over 75%, for the *thumb/pinky item* in English, the Production model predicts 100% of participants selecting the pinky finger over the thumb. Otherwise, the model predictions fall in line with all empirical results for the comprehension experiments in Spanish and English.

Figure 10 plots the rate at which listeners chose the image on the right along the x-axis against the probability assigned to the image on the right by each model on the y-axis. A perfect model would assign probability at the exact same rate as actual production. The R^2 for the Complexity model is only 30.6%. This means that the Complexity model accounts for 30.6% of the variation present

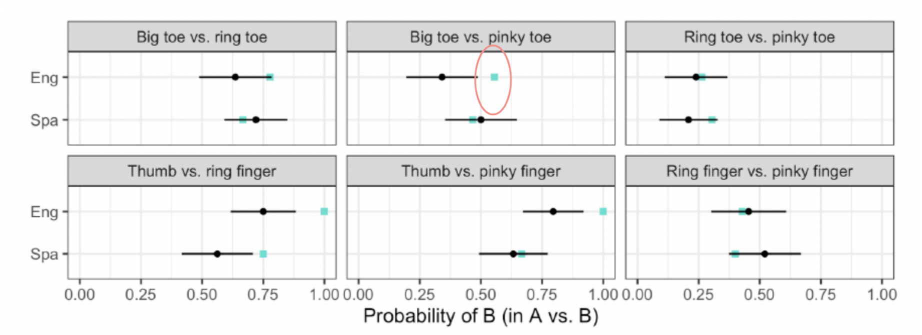


Figure 9: Production Model predictions (square) plotted against comprehension results; Inaccurate model predictions are circled in red

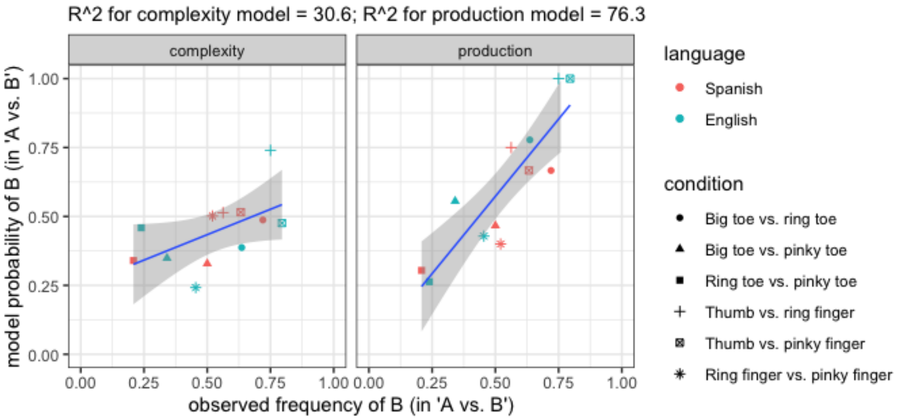


Figure 10: Comparison of Model Results

in the comprehension studies. The R^2 for the Production model, in contrast, is 76.3%, which is to say that the Production model accounts for 76.3% of the variance in the data. There is a stark contrast in the explanatory power of each model. This shows that listeners have a good mental model of speakers, and that their mental model is not purely complexity-based. In fact, the comparison of these model results suggests that speakers are considering prevalence *over* complexity, since the prevalence-based Speaker model does not include a cost parameter.

7 Discussion

Figure 11 summarizes the scalar implicatures found in my comprehension results. The highlighted digits represent the target digit. Arrows point toward the image that speakers preferred, suggesting an implicature that negates the digit on the other side of the arrow. Dotted lines indicate that speakers did not calculate an implicature when presented with the two digits on either side of the dotted line. As Figure 11 shows, Spanish and English speakers do differ with respect to the scalar implicature associated with *finger* in accordance with the prevalence of the words for 'thumb', 'ring finger', and 'pinky finger'. In addition, it seems that, contrary to previous theories (Horn, 1984, 2000; Geurts, 2011), English and Spanish speakers do calculate an implicature associated with *toe*. These findings will be discussed in further detail below.

The results presented above suggest that Horn was correct – if English had *pollex* (like Spanish *pulgar*), the *finger* $\rightsquigarrow$ 'not thumb' implicature would disappear. Additionally, Geurts's predictions were also correct. The decision to include or exclude an alternative in implicature calculation is influenced by how "colloquial" the alternative is. This is exactly in accordance with the prediction that *dedo* 'finger' does not narrow in opposition to *pulgar* 'thumb' due to the lower prevalence of *pulgar*. These results line up with the production results extremely well. *Thumb* is such a prevalent alternative in English that it was produced by all English speakers. In stark contrast, Spanish speakers produced seven different utterances *in addition* to *pulgar* when presented with an image of the thumb.

When we compare *ring finger* and *pinky finger* in the comprehension results, there is no preference for either image from English and Spanish speakers. This result echoes the pattern found in the production results for the ring finger item and the pinky finger item, since they both elicit similar rates of specificity in production. The production results show that the single-word *pinky* is available to speakers to refer to the pinky finger, but it is not as prevalent. From a complexity perspective the comprehension results are puzzling – why doesn't *finger* (or *dedo*) imply 'not pinky'? After all, both *pinky* and *meñique* are single words. Horn (2000) says "What is crucial is the status of *thumb* as opposed to *pinky* as a viable lexicalized alternative to *finger*" (p. 308). That is to say, Horn believes *pinky* must not be a *viable* lexicalized alternative, while *thumb* is. However, this conclusion is problematic. On what grounds is *pinky* excluded from the class of viable lexicalized alternatives? This is especially troubling, considering *pinky* and *meñique* were the most frequent responses from participants in the production study.

I posit that there exists a small probability that *pinky*, as opposed to *pinky finger*, is not in the lexicon due to dialectal variation, and speakers are acutely aware of this. Initially, I suspected that *pinky finger* had such a high rate of production due to speakers' desire to avoid ambiguity – that *pinky* could potentially refer to the pinky finger or the pinky toe. However, the production results suggest that *pinky* is not an available alternative to refer to the pinky toe: zero participants produced *pinky* in reference to the pinky toe item. To

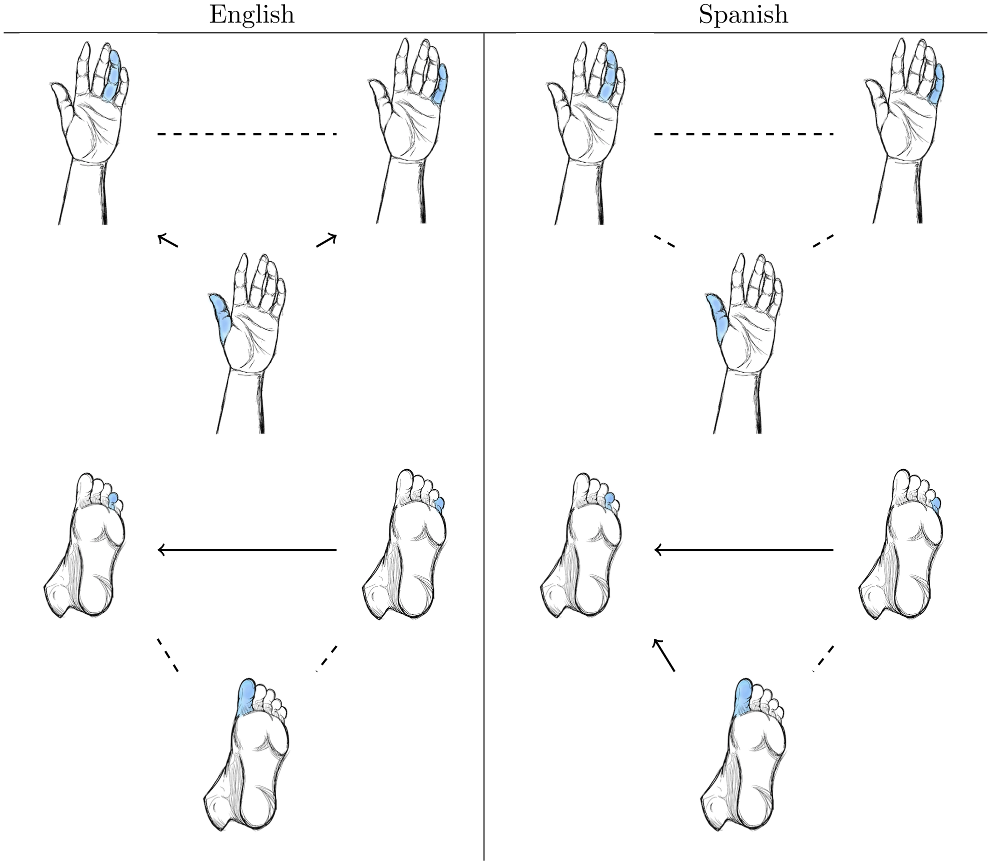


Figure 11: Summary of English and Spanish Implicatures; Dashed lines indicate no implicature was calculated when participants were presented with the two images on each side of the dashed line; solid arrows indicate an implicature was calculated in the direction of the arrow – when participants were presented with the images on either side of the arrow, they calculated the implicature ‘not X’, where *X* stands for the digit the arrow is pointing away from

further support this, native speakers that I have consulted informally tend to say that ‘She has a tattoo on her pinky’ is false when referring to the pinky toe. This supports the conclusion that the prevalence of the term *pinky finger* has less to do with ambiguity and more to do with uncertainty about whether *pinky* is in the lexicon. The risk of ambiguity could be an explanation for why some Spanish speakers utilized *dedo de la mano* ‘finger’ and *dedo del pie* ‘toe’ over the ambiguous form *dedo* ‘digit’ in the production results. Including the words *de la mano* or *del pie* to avoid ambiguity seemingly lowers the prevalence of the specific terms available to speakers.

In addition, it is not clear that prevalence is only relevant for *lexicalized* alternatives, that is to say alternatives that are present in the lexicon, as Horn (2000) puts it. When a complex alternative is more prevalent than another equally complex alternative, one can see an implicature here as well. Horn excludes *big toe* as a lexicalized alternative, since no implicature arises for *toe*. Thus, by his criteria we would not want to consider *pinky toe* and *ring toe* to be lexicalized alternatives either. However, the results from the comprehension study show that participants calculated an implicature for the pinky toe/ring toe item. That is, in the context of two images — a tattooed pinky toe and a tattooed ring toe — English and Spanish listeners calculate an implicature that *toe* implies ‘not pinky toe’ given the utterance *she has a tattoo on her toe*.

What these results suggest is that whether or not an implicature arises has more to do with *what a speaker would say* and less to do with complexity *per se*. Additionally, the production results show a general tendency to use a more specific description with *pinky toe*. Since it is substantially more likely that speakers will use a general description for the ring toe than for the pinky toe, listeners possess this metalinguistic awareness that allows them to calculate an implicature.

The comprehension results described above point toward a model of implicature that is grounded more in production probability – what speakers say – than in complexity. This conclusion is strengthened by the accuracy of the prevalence-based model when compared to the purely complexity-based model. As previously discussed, the Production model was much more accurate at predicting scalar implicature calculations than the Complexity model. When the listener is only considering the prevalence of an utterance, which is the case for the Production model, the model is much better at predicting where scalar implicatures will be calculated.

But these conclusions lead to further questions. We know that listeners are Bayesian in nature, and they have a good model of speaker production. However, the Production model does not attempt to predict what the speakers are going to do, production-wise. Because the Production model contained the actual production data from my experiments, no predictions were necessary. The Complexity model *does* attempt to predict speaker production, but, as the results show, it does not do so very accurately.

Another question surrounds the operationalization of complexity. For the purposes of this study, I considered *pulgar* to be monomorphemic, which informed my decision to measure complexity at the word level. It is possible that measuring complexity using a metric other than word count (e.g. syllable or morpheme) could lead to an increase in model accuracy.

A third question that arises when examining the experimental results is the significance of the dispersion in the production data. There is a clear contrast between the number of alternatives apparently available for English speakers and Spanish speakers. Spanish speakers present many more unique utterances than English speakers, and I am unsure of the effects, if any, these differences have on scalar implicature calculation. Does the larger amount of dispersion in

Spanish production data correlate with the lack of scalar implicature in Spanish for all digits in the hand? It is possible that listeners have a perfect understanding of speaker production, and higher dispersion weakens the prevalence of all possible alternatives. This could explain the absence of scalar implicatures for the fingers, but this conclusion does not align with the fact that Spanish speakers *do* calculate scalar implicatures for the big toe/ring toe item and the ring toe/pinky toe item.

8 Conclusions

The results outlined above suggest that Spanish and English speakers do differ with respect to the scalar implicatures associated with *finger* in accordance with the prevalence of the words for 'thumb', 'ring finger' and 'pinky finger'. My empirical results support the idea that differences across languages in the implicatures associated with general terms are closely tied to differences in production probabilities for more specific terms. Since *pulgar* 'thumb' is not as prevalent in Spanish as *thumb* is in English, it is not available in the set of alternatives to *finger*, which is why speakers do not calculate an implicature. My model results further support the conclusion that alternatives are constrained based on prevalence: the Production model significantly outperforms the Complexity model. While complexity does assist in determining the set of alternatives present for speakers, it is not as explanatory as full awareness of what speakers actually produce. Crucially, these findings go against structural theories, like Katzir (2007) and Horn (2000), that constrain the set of alternatives based on complexity alone. They support pragmatic frameworks, like Geurts (2011) and RSA (Frank & Goodman, 2012) that propose constraints on the set of alternatives based on prevalence.

Further research should examine the scalar implicatures that participants seem to be calculating in English and Spanish for the toes. To my knowledge, these findings are the first of their kind. This suggests a possible shift in the prevalence of specific terms for the toes, since the critical example for this paper stemmed from an asymmetry between *finger* and *toe* – where the implicature *finger* $\rightsquigarrow$ 'not thumb' arrises, while the implicature *toe* $\rightsquigarrow$ 'not big toe' does not. Another area for future research surrounds the operationalization of complexity. My findings hold true only if Spanish *pulgar* 'thumb' is equal in complexity to *thumb*. It is possible that this is not the case, and that speakers store complexity in terms of a measure of something other than word count (possibly morpheme or syllable).

Ultimately, this paper presents a novel approach using cross-linguistic comparison to investigate how speakers constrain the set of alternatives when calculating scalar implicatures. My findings point away from traditional Gricean and Neo-Gricean methods for imposing restrictions on alternatives. Instead, my findings support a Bayesian approach to constraining the set of alternatives that is centered around the production probability of alternatives. Listeners have a keen awareness of prevalence, and this informs their scalar implicature calculation.

References

Andersen, Gisle. 2017. Pragmatic borrowing of discourse items: a challenge for cross-linguistic pragmatics. In *Proceedings of the workshop on cross-linguistic pragmatics at the leibniz-centre general linguistics*, .

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Atlas, Jay & Stephen Levinson. 1981. It-clefts, informativeness, and logical form: radical pragmatics (revised standard version). In Peter Cole (ed.), *Radical pragmatics*, 1–61. New York: Academic Press.

Bergen, Leon, Noah D. Goodman & Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society*, 120–125. Cognitive Science Society.

Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics & Pragmatics* 9(20). 1–83.

Blutner, R. 2000. Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17(3). 189.

Blutner, Reinhard. 1998. Lexical pragmatics. *Journal of Semantics* 15(2). 115–162.

Breheny, Richard, Nathan Klinedinst, Jacopo Romoli & Yasutada Sudo. 2018. The symmetry problem: current theories and prospects. *Natural Language Semantics* 26(2). 85–110.

Chierchia, Gennaro. 2004a. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (ed.), *Structures and beyond*, Oxford: Oxford University Press.

Chierchia, Gennaro. 2004b. A semantics for unaccusatives and its syntactic consequences. In Artemis Alexiadou, Elena Anagnostopoulou & Martin Everaert (eds.), *The unaccusativity puzzle: Explorations of the syntax-lexicon interface*, 22–59. Oxford University Press.

Chierchia, Gennaro. 2005. Broaden up your view. Implicatures of domain widening and the 'logicality' of language. Ms. University of Milan.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2008. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, Berlin: Mouton de Gruyter.

von Fintel, Kai & Danny Fox. 2002. Classnotes for 24:954: Pragmatics in linguistic theory.

Fox, Danny. 2007. Free choice disjunction and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. Basingstoke: Palgrave Macmillan.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998.

Franke, Michael & Gerhard Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft* 35(1). 3–44.

Gazdar, Gerald. 1979. A solution to the projection problem. In Choon-Kyu Oh & David Dineen (eds.), *Syntax and semantics 11: Presupposition*, 57–89. New York: Academic Press.

Geurts, Bart. 2005. Entertaining alternatives: Disjunctions as modals. *Natural Language Semantics* 13(4). 383–410.

Geurts, Bart. 2011. *Quantity implicatures*. Cambridge: Cambridge University Press.

Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11). 818–829.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.

Grice, H. Paul. 1981. Presupposition and conversational implicature. In P. Cole (ed.), *Radical pragmatics*, 183–98. New York: Academic Press.

Grice, Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics*, vol. 3, 41–58. New York: Academic Press.

Horn, Laurence. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Washington, DC: Georgetown University Press.

Horn, Laurence. 1991. Duplex negatio affirmat... the economy of double negation. *Chicago Linguistics Society* 27. 80–106.

Horn, Laurence R. 1972. On the semantic properties of logical operators in English. Doctoral dissertation, University of California, Los Angeles.

Horn, Lawrence R. 2000. From *if* to *iff*: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics* 32(3). 289–326.

Jäger, Gerhard. 2000. Some notes on the formal properties of bidirectional optimality theory. *ZAS Papers in Linguistics* .

Jäger, Gerhard. 2012. Game theory in semantics and pragmatics. *in Maeinborn et al. (2012)* 2487–2425.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690.

Kiparsky, Paul. 1982. Word formation and the lexicon. In F. Ingemann (ed.), *Proceedings of the 1982 Mid-America Linguistic Conference*, 3–32. University of Kansas, Lawrence, KS.

Krifka, Manfred. 1989. Nominal reference, temporal constitution and quantification in event semantics. In Renate Bartsch, Johan van Benthem & Peter van Emde Boas (eds.), *Semantics and contextual expression*, 75–115. Dordrecht, Netherlands: Foris.

Levinson, Stephen C. 2000. *Pragmatik: Neu übersezt von martina wiese*. Tübingen: Max Niemeyer.

Matsumoto, Yo. 1995. The conversational condition on horn scales. *Linguistics and philosophy* 18(1). 21–60.

Matthewson, Lisa. 2006. Temporal semantics in a superficially tenseless language. *Linguistics and Philosophy* 29. 673–713.

McCawley, J.D. 1978. Conversational implicature and the lexicon. In Peter Cole (ed.), *Syntax and semantics, volume 9: Pragmatics*, 245–259. Academic Press.

Renans, Agata. 2012. Projective behaviour of *nur:* quantitative experimental research. In *Logic, language and meaning: 18th amsterdam colloquium, amsterdam , the netherlands, december 19-21, 2011, revised selected papers*, 190–199. Berlin: Springer.

van Rooij, Robert & Katrin Schulz. 2004. Exhaustive interpretation of complex sentences. *Journal of Logic, Language, and Information* 13. 491–519.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.

Sperber, Dan & Deirdre Wilson. 1986. *Relevance*. Cambridge, Massachusetts: Harvard University Press.

Stateva, Penka, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy & Anne Colette Reboul. 2019. Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology* 10. 957. doi:10.3389/fpsyg.2019.00957. `https://www.frontiersin.org/article/10.3389/fpsyg.2019.00957`.

Swanson, Eric. 2010. Structurally defined alternatives and lexicalizations of xor. *Linguistics and Philosophy* 33(1). 31–36.

Teodorescu, Viorica Alexandra. 2009. *Modification in the noun phrase: the syntax, semantics, and pragmatics of adjectives and superlatives*: University of Texas at Austin dissertation.

van Rooij, Robert. 2003. Negative polarity items in questions: strength as relevance. *Journal of Semantics* 20. 239–73.

Yuan, Arianna, Will Monroe, Yu Bai & Nate Kushman. 2018. Understanding the rational speech act model. In *The 40th annual meeting of the cognitive science society*, .

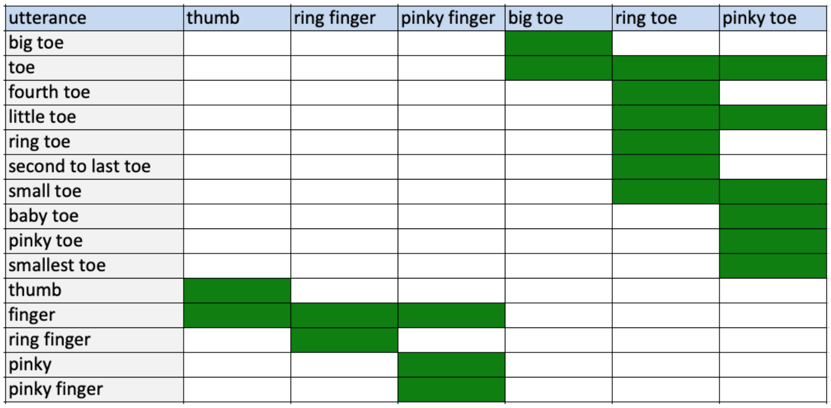
Zimmermann, Ede. 2000. Free choice disjunction and epistemic possibility. *Natural Language Semantics* 8. 255–290.

Appendices

A Literal Meanings

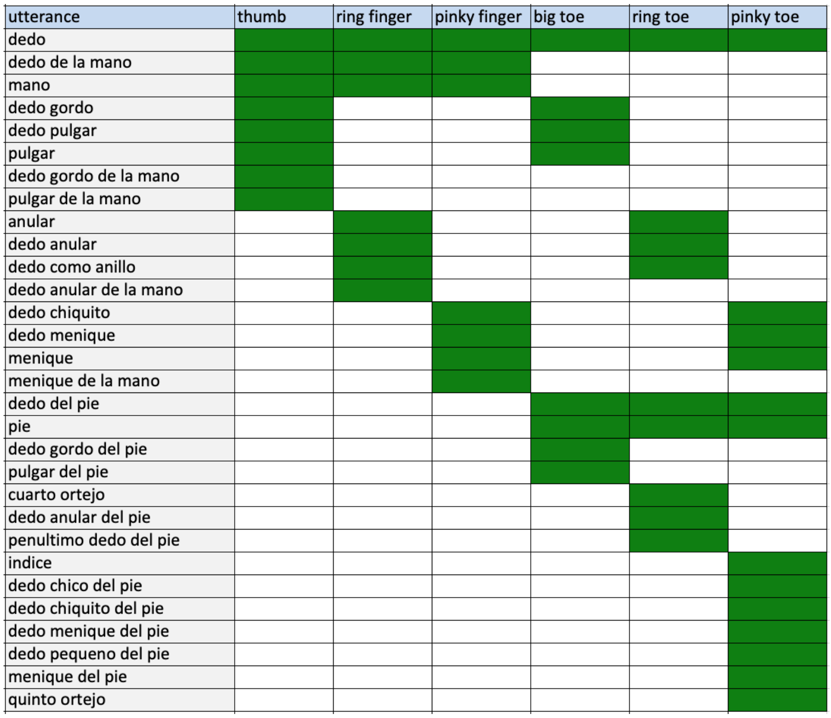
Below are the hand-coded literal meanings in English and Spanish for all utterances collected from the production studies. Green boxes represent accurate utterance/state pairings, where the utterance corresponds to the underlying state.

One interesting observation is that Spanish speakers presented a much higher variety of responses, suggesting that Spanish speakers have a larger set of alternatives available to them. This, therefore, led to a larger set of literal meanings for my models. While this is outside of the scope of this paper, future work could investigate the effects of this disparity.



utterance	thumb	ring finger	pinky finger	big toe	ring toe	pinky toe
big toe				🟩		
toe				🟩	🟩	🟩
fourth toe					🟩	
little toe					🟩	🟩
ring toe					🟩	
second to last toe					🟩	
small toe					🟩	🟩
baby toe						🟩
pinky toe						🟩
smallest toe						🟩
thumb	🟩					
finger	🟩	🟩	🟩			
ring finger		🟩				
pinky			🟩			
pinky finger			🟩			

Figure 12: **English Literal Meanings**



utterance	thumb	ring finger	pinky finger	big toe	ring toe	pinky toe
dedo	🟩	🟩	🟩	🟩	🟩	🟩
dedo de la mano	🟩	🟩	🟩			
mano	🟩	🟩	🟩			
dedo gordo	🟩			🟩		
dedo pulgar	🟩			🟩		
pulgar	🟩			🟩		
dedo gordo de la mano	🟩					
pulgar de la mano	🟩					
anular		🟩			🟩	
dedo anular		🟩			🟩	
dedo como anillo		🟩			🟩	
dedo anular de la mano		🟩				
dedo chiquito			🟩			🟩
dedo menique			🟩			🟩
menique			🟩			🟩
menique de la mano			🟩			
dedo del pie				🟩	🟩	🟩
pie				🟩	🟩	🟩
dedo gordo del pie				🟩		
pulgar del pie				🟩		
cuarto ortejo					🟩	
dedo anular del pie					🟩	
penultimo dedo del pie					🟩	
indice						🟩
dedo chico del pie						🟩
dedo chiquito del pie						🟩
dedo menique del pie						🟩
dedo pequeno del pie						🟩
menique del pie						🟩
quinto ortejo						🟩

Figure 13: **Spanish Literal Meanings**